Linguistic and Prosodic Emotion Recognition

Tim Polzehl[1], Florian Metze[2], Alexander Schmitt[3]

[1] *Quality and Usability Lab / Deutsche Telekom Laboratories, 10587 Berlin, Germany, Email: tim.polzehl@telekom.de*

[2] *LTI, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A., Email: fmetze@cs.cmu.edu*

[3] *Dialogue Systems Group, 89081 Ulm, Germany, Email: alexander.schmitt@uni-ulm.de*

Abstract

This paper compares the performance of linguistic and acoustic/ prosodic features for emotion recognition by evaluating them on two speech databases collected using deployed customer care IVR systems. In our anger recognition task, the target is to recognize emotional user states, using either acoustic properties of the speech signal, or a transcription of what was said. Robust processing of speech will require the examination both sources, and an understanding of how these sources behave on different corpora. We present dependencies of both approaches and outlines future improvements. Given the present database design, acoustic/ prosodic modeling clearly outperforms linguistic modeling.

Introduction

Emotion detection in Interactive Voice Response (IVR) Dialog systems can be used to monitor quality of service or to adapt emphatic dialog strategies [1, 2]. Especially anger detection can deliver useful information to both the customer and the carrier of IVR platforms. It indicates potentially problematic turns or slots to the carrier so he can monitor and refine the system. It can further serve as trigger to switch between tailored dialog strategies for emotional conditions to better react to the user's behavior [3]. Some carriers have also been experimenting with re-routing the customers to the assistance of a human operator when problems occur. Problems and uncertainties arise from the imbalance in complexity between human computer interaction and models trained for these interactions. The difficulty is to capture the various and divers human expression patterns that convey emotional information by automated measurements.

On the one hand this paper presents results from acoustic and prosodic anger recognition, i.e. we examine expressive patterns that are based on vocal intonation. Applying our system from [4] we capture these expressions extracting low-level audio descriptors, e.g. pitch, loudness, MFCC, spectrals, formants and intensity. After extraction statistics are applied to the descriptors. These statistics serve as model parameters.

On the other hand we apply linguistic feature classification, i.e. we analyze the words the users speak. Modeling mutual information of the two distributions at hand, i.e. the probability of emotions and the probability of emotions given certain words we calculate the Emotional Salience as described in [5]. Eventually, the turn-wise accumulated salience scores serve as model parameters.

Databases

Both our database consists of 'realistic' speech, i.e. they have background noise, recordings include cross- and off-talk. In principle, the databases comprise speech from the same domain, i.e. Internet- and telephone-related services and troubleshooting.

The German IVR database roughly captures 21 hours recordings from a German voice portal. The annotated labels can be collapsed to a binary division between *Anger* and *Non-Anger* utterances [6]. Inter labeler agreement results in $\kappa = 0.52$. Our experimental subset contains 1951 Anger turns and 2804 Non-Anger turns which roughly corresponds to a 40/60 split of anger/non-anger distribution. The average turn length after cleaning out initial and final pauses results in 1.8 seconds. Mirroring the same class distribution we extract a subset of 1560 *Non-Anger* and 1012 *Anger* turns form the English IVR database of overhaul 10h of recordings. The inter labeler agreement results in $\kappa = 0.63$, which also resembles moderate agreement. The average turn length after cleaning pauses is approx. 0.8 seconds.

Acoustic Anger Classification

The audio descriptors can be sub-divided into 7 groups. These extracts base on *pitch, loudness, MFCC, spectrals, formants, intensity* and *other* features.

We extract *pitch* by autocorrelation. After converting pitch into the semitone domain we apply piecewise cubic interpolation and smoothing by local regression using weighted linear least squares. We extract 16 Mel frequency cepstral coefficients, $MFCC$. Further descriptors from the *spectral* domain are the center of spectral mass gravity, the 95% spectral energy roll-off point and the the magnitude of spectral change over time. Using linear predictive coding (LPC) we extract 5 $formant$ frequencies and estimate their bandwidths. An integration of spectral Barc filtering yields a *loudness* estimate. Taken from the time domain we extract *intensity* in decibel. Referred to as *other* features we calculate, e.g., the Harmonics-to-Noise Ratio (HNR), the correlation between pitch and intensity, the Zero-Crossing-Rate, and the relation of pitched and non-pitched speech segments as individual features. All descriptors are extracted using 10ms frame shift.

After extraction we calculate statistics such as means, moments of first to fourth order, extrema and ranges from the respective descriptors. Special statistics are applied to certain descriptors, e.g. pitch, loudness and intensity

are further processed by a discrete cosine transformation (DCT) to estimate the spectral composition. Both our databases comprise very short utterances. We presume that every turn is a short utterance of one prosodic entity and model all features on turn level. In order to exploit the temporal behavior we append derivatives and calculate statistics on them alike. Segmenting the audio signal into voiced, unvoiced and silence frames we also compute quotients of the features from these segments. A more detailed description can be found in [7].

Linguistic Anger Classification

Linguistic features model the information given by the words the users choose. Departing from the relation of class specific word usage and prior class distribution knowledge Lee calculates the *Emotional Salience* of a word using the mutual information between the probability of words and the probability of emotion classes [5]. Let k be the number of classes, $P(\epsilon)$ the prior probability of an emotion and $P(\epsilon|w)$ the posterior probability that an utterance containing a word $w \in W$ implies an emotion class $\epsilon \in E$. He defines the Emotional Salience of a word as:

$$MI(E; W = w) = \sum_{j=1}^{k} P(\epsilon_j|w) \cdot log \frac{P(\epsilon_j|w)}{P(\epsilon_j)} \qquad (1)$$

On turn level, he sums up the word- and class-specific salience values and decides for the class of maximum accumulated score.

Results

Table 1 shows the classification results of the linguistic and acoustic features for the databases. Scores are obtained by Support Vector Machine (SVM) classification using a linear kernel function. The high number of acoustic features are reduced by applying a ranking filter, i.e. Information Gain Ratio (IGR). Feature spaces for both databases are determined separately by increasingly appending top-ranked features until the classification performance reaches a maximum. All results and system parameter tuning steps base on 10 fold cross-validation. Turns of unknown words, e.g. in the cross-validation test splits, are assigned to the majority class. Classification success is measured using overall accuracy and, to give a class distribution independent estimate, as f1 score.

Looking at the performance figures, the acoustic features generally work better than the linguistic features. The overall acoustic feature performance does not vary significantly when comparing the databases, i.e. the features seem to be robust in terms of the database conditions and languages. The linguistic feature performance is generally much lower. Note, that constant voting for the majority class would result in an accuracy of 60% and an f1 of approx. 40% already. The relative low F-measure for the Anger class indicates insufficient word models. This can be due to database design, e.g. the level of system initiative which is a menu-based directed dialog in the present case of the English IVR. If the users responses

Table 1: Acoustic and Linguistic Classification Scores

Feature	German IVR	English IVR
Linguistics		
f1	63.6%	57.9%
accuracy	66.1%	64.7%
F_A	54.5%	41.1%
F_{NA}	72.8%	74.7%
Acoustics		
f1	77.2%	77.0%
accuracy	79.1%	78.4%
F_A	67.6%	69.8%
F_{NA}	86.7%	84.1%

are frequently restricted, e.g. 'yes, no, continue,' etc., the emotional expression is predominantly transmitted by means of intonation. Modeling acoustic/prosodic patterns seems consistently more promising, since the standard deviation from the different splits varies less than than 4% only.

Future work will examine, how factors as dialog design, turn length or word perplexities can be used to predict linguistic modeling performance. Presumably, increasing turn lengths contributes to better linguistic performance as word models can then be estimated more robustly, also by using n-grams. On the other hand, turn-wise acoustic statistics might then be distorted by overlay of other contiguous prosodic entities that might not always be intended to carry emotional meaning. Intelligent decision fusion, e.g. by confidences, will be in the focus of future research.

References

[1] Yacoub, S., Simske, S., Lin, X., Burns, J.: Recognition of Emotions in Interactive Voice Response Systems. Eurospeech. 2003

[2] Shafran, I., Riley, M., Mohri, M.: Voice Signatures. IEEE ASRU. 2003

[3] Metze, F., Metze, Englert, R., Bub, U., Burkhardt, F., Stegmann, J.: Getting Closer: Tailored Human-Computer Speech Dialog. Universal Access in the Information Society. Springer. 2008

[4] Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., Metze, F.: Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. Interspeech. 2009

[5] Lee, C. M., Narayanan, S. S.: Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing. 2005

[6] Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R.: Detecting Real Life Anger. ICASSP. 2009

[7] Minker, W., Lee, G. G., Mariani, J., Nakamura, S.: Salient Features for Anger Recognition in German and English. Spoken Dialogue Systems Technology and Design. Springer, Boston (USA). to appear 2010